JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125 VOL 7, ISSUE 19, 2020

A REVIEW ON MACHINE LEARNING (FEATURE SELECTION, CLASSIFICATION AND CLUSTERING) APPROACHES OF BIG DATA MINING IN DIFFERENT AREA OF RESEARCH

Neeraj[1], Narender Kumar[2] , Vineet Kumar Maurya[3]

[1,2]Department of Computer Science and Engineering, H. N. B. Garhwal University (A Central University) Srinagar Garhwal, Uttarakhand, India
[3]Department of Botany and Microbiology, H. N. B. Garhwal Univeristy, (A Central University), Srinagar Garhwal, Uttarkhand, India

E-mail: vineetkm2000@gmail.com, narenrawal@gmail.com

Received: 14 March 2020 Revised and Accepted: 8 July 2020

ABSTRACT: Today's age is the age of data, where a huge amount of data is being generated world-wide. This huge volume of data, called 'big data', has no meaning until the proper information is extracted from it. A small size data, with limited number of dimensions can be analyzed using simple computer programs like MS-excel sheets, but a huge, complex and multidimensional data (big data) cannot be analyzed using simple computational methods. Hence, machine learning approaches of data mining is required, which includes process of feature selection, classification and clustering. All these steps are complex in themselves and require special algorithms for each. A survey of theoretical insights of different methods used for feature selections (FS), clustering and classification is presented in this review. Application of machine learning in marketing, library, climate, crime and biological sciences, etc. is also briefly mentioned in this review. I was also observed that that there is no any perfect method that can analyze different type of big datasets with equal efficiency and accuracy. Moreover, hybrid algorithms (hybrid filter, hybrid wrapper, hybrid Evolutionary based algorithms) and modified algorithms are preferred for big data mining due to their better performance over ensemble approach or single algorithms. Use of more than one methods or modified or hybrid methods for better prediction accuracy is a good choice. Despite of many combinations and trial still there is no single algorithms that can be universally applied for big data arising in different fields, hence newer algorithms or better amalgamations of different algorithms are still required, for analysis of big data. Absence of any perfect algorithms and continuous generation of data has a great scope for computational scientist and programmers.

KEY WORDS: Big data, Clustering, Feature selection, Classification, SVM, Decision tree, Naïve Bayes, Genetic Algorithms

I. INTRODUCTION

Present age is the age of 'big data' [1] where a huge volume of data is already available and the same is being generated every day. In order to retrieve meaningful information the data must be processed or mined systematically with clearly defined goals[2]. The processed information leads to knowledge building, that's why some time data mining methods are called 'Knowledge mining' or 'knowledge discovery [3]. A large volume of data is being generated every day in every field and extraction of meaningful information out of it is a fundamental need. Market analysis [4], psychology analysis of consumers and students, medical [5] and pharmacological data analysis [6], biological data analysis[7], climate change, accidents and crime analysis [8] are the various fields where data mining is highly demanded. Data mining is umbrella term and can be used in different context according to demand. Extracting information from unorganized data, (news articles, share market, demand and supply data from market, research papers and survey etc.) is also a type of data mining called text mining [9].Likewise mining of organized data in form of databases (biological databases, chemical database, crime database, library databases, etc.) is also known as data mining [10]. There are different variations of data [11] such as simple/complex, organized/unorganized small/large, numerical/categorical, binary/multiple and one-dimensional/multi-dimensional, etc. Microarray datasets, chemical databases, library databases are the perfect example of organized, large, discrete, numerical, nominal, multidimensional data, organized into binary or multiple classes [12]. Traditional methods of data mining (dealing with single information at a time) can deal simple, organized and small sized data, but are not amenable for large, complex,

ISSN- 2394-5125 VOL 7, ISSUE 19, 2020

multidimensional data; hence advance computational methods of data mining are required to divulge knowledge from such type of data.

II.DATA MINING AND MACHINE LEARNING

Data mining doesn't mean digging or mining of data, but the extraction of hidden information (pattern, interactions, correlations, groups, features, outliers, etc.) from it. Along with experimental part for data collection, mining of data is also inevitable in all types of research. Data mining technologies save the time and reduce the chances of manual error, because big-data cannot be handled efficiently and accurately by mere human mind. Analysis of big-data using computers required a good amalgamation of database management and retrieval systems, statistics and programming language. All these are essential components of big-data mining. There are three fundamental steps of big-data mining; data pre-processing, machine learning (ML) and artificial intelligence (AI) [13]. In data pre-processing, missing values are imputed and non-significant dimensions of data are eliminated. Machine leaning is the heart of data mining where various statistical procedure and applied on big-data sets to find hidden facts in it. Based upon results given by ML, artificial intelligence infers some conclusions about data. This conclusion is further to be tested using real human intelligence. Amongst three steps of big-data mining, the ML approaches are discussed briefly in this article, which includes three processes; feature selection, classification and clustering.

III.FEATURE SELECTION

A complex, multidimensional data usually have four types of features; (i) high weighted features (most relevant and non-redundant) (ii) medium weight features (weakly relevant but non-redundant features) (iii) less-weighted features (weakly relevant and redundant features), and (iv) zero-weighted features (completely irrelevant or noise), [14]. Feature selection (FS), also known as variable or attribute selection, is process of eliminating less-weighted and irrelevant features and picking most relevant features of the data. This improves the prediction accuracy, validity of extracted information and reduces computational cost, processing time. There would be 2^n-1 attributes in a data with 'n' number of dimensions and analysis of such data could be computationally infeasible task if 'n' is too high. FS is used to reduce data dimensions and eliminate 'curse of dimensionality' of big data by selecting significant features [14]. Literature survey confirms that 'classifications' performed with FS outperforms the 'classification' done without FS, in term of accuracy and speed [15]. Algorithms used for FS (FSAs) are can be supervised (labels are provided to each data), semi-supervised (labels are provided to part of data) and unsupervised (labels are not provided to any part of data). On the basis of methods applied for feature searching, FSAs can be grouped into four types: Filter, Wrapper, Embedded methods and Hybrid methods [16], which are described here.

A. Filter method

Filter methods evaluates the features on four criteria namely information theory, dependency, consistency and distance. Filter methods select most discriminative features amongst all, using inherent characteristics of data without applying any algorithms. Filters calculate the degree of correlation between selected feature and output class label. These degree of correlation (correlation scores) are used for ranking of features and the high ranking features are selected. Filter methods are faster and less computer intensive. Different types of filters are; statistics based features [Chi-square (Witten and Frank 2002), t-test, F-statistics, ANOVA, Principle component analysis (PCA), [17], Euclidean distance (ED) [18], Bhattacharya Distance [19] , Cosine similarity [18] and correlation based feature selection (CBFS) [20]], mutual information based [Information gain (IG) [21], Gain ratio (GR) (Witten and Frank 2002) [20], Symmetrical uncertainty (SU) [22], entropy and pure continuity] and others like Minimum redundancy maximum relevance (mRMR)[23], Inconsisten cy criterion (IC) [24], Relief (ReliefF, ReliefC etc.) [25] Filter can be univariate (evaluate only one feature at a time) or multivariate (evaluate subset of features at a time). $\chi 2$ statistics, t-test, F-statistic, ED, IG, GR , LS and ED are the example of univariate filters. An example of univariate filters based FSAs are ReliefC and RelifF, which are used for clustering. Spectral feature selection (SPEC) [26] is the example of univariate filters based FSA used in both, classification and clustering. In multivariate filters based FSAs, CBFS, mRmR, IC, Fast correlation-based feature selection (FCBF) [27], Markov blanket filter (MBF) [28]) are used for classification. While Feature selection for sparse clustering (FSSC), Localized Feature Selection Based on Scatter Separability (LFSBSS), Multi-Cluster Feature Selection (MCFS) [29] and feature weighting k-means [30] are example of multivariate filter based FSAs used for clustering.

B. Wrapper method

In wrapper methods [31], instead of individual features, subsets of most relevant features are selected and evaluated one by one, using classification accuracy as fitness function. These are close-loop methods [32] and

ISSN- 2394-5125 VOL 7, ISSUE 19, 2020

used in both, clustering and classification algorithms. Forward selection, backward selection and recursive feature elimination are the approaches used in wrapper methods. Due to repetitive evaluation, Wrapper methods are much slower and computer intensive than filter methods. Wrappers can be deterministic or randomized. Sequential forward selection (SFS), Sequential backward elimination (SBE) Plus-L Minus-R selection (LRS) [33], Smart Beam search (SBS) [33] algorithms are used with deterministic wrapper, while Simulated annealing (SA), Randomized hill climbing, Genetic algorithms (GA), Estimation of distribution (ED) are used with randomized wrapper based FSAs.

C. Embedded method

Embedded methods [31], incorporates FS as an integral part of clustering or clustering algorithms and FS is performed during their execution. As name indicates, these methods are embedded in the algorithm either as its normal or extended functionality, and use a unique 'sparsity regularization algorithms'[34] such as LASSO [35], Ridge Regression, Elastic Net (RREN), which makes the weight of some features zero. Decision tree (DT), Random forest (RF), Artificial neural network (ANN), Naïve Bayes (NB) and Support vector machine (SVM) are some classification algorithms, with which embedded methods of FS are used.

D. Hybrid method

Hybrid methods are either a combination of more than one FS methods [36], or a modified version of already existing FSAs. Unlike ensemble methods, hybrid methods use different FS methods sequentially on entire dataset. Hybrid methods combine high efficiency of filter methods, high accuracy of wrapper methods and reduce computational complexities. Hybrid methods first use filter methods to reduce the dimensions of the data and then wrapper find the best candidate subset. Common hybrid methods are fuzzy random forest based FS [37], hybrid genetic algorithms e.g. SRPSO [38], hybrid ant colony optimization (WFACOFS) 201[39], or mixed gravitational search algorithm [40]. MGRFE [41], TLBOSA [42] and TLBGOSA [43] are examples of hybrid wrapper algorithms, while Multiple Multiclass Artificial Bee Colony' (MMABC) algorithm [44], MOCEPO [45] are few example of modified algorithms.

After the relevant features are identified using FSAs, next step of data mining is clustering or classification. These are the fundamental tasks of data mining, involving unsupervised learning and supervised learning, respectively [46]. Different approaches and algorithms for clustering and classifications are mentioned here. Due to space constraint, only their introductions and important features are given, without digging into technical details. Although different approaches and algorithms are listed under clustering and classification subsections, but these approaches and algorithms are not limited to those subsections only. There are various examples in which different approaches have been used for both clustering and classification, both.

IV.CLUSTERING:

Cluster is a collection of similar features in a single subset, which can be treated as a implicit class. Aim of clustering is to make meaningful and coherent subsets of similar features out of given datasets. In clustering process similar features are kept in same group while dissimilar features are always kept in other group [47]. Clustering can be top-down in which entire dataset is considered as single cluster on the base of any common feature, further clusters are made from it by separating a smaller clustering having lesser similarity. Bottom-up approach of clustering consider each feature as one cluster-point and size of clusters are increased by incorporation additional similar features to different cluster points. Clustering can be hard (non-overlapping), based on actual values or while in fuzzy (overlapping), based on probability [48]. Clustering is also known as data (descriptor) segmentation or unsupervised automatic classification, because it partitions large datasets into smaller groups according to their similarity. Clustering can also be used for outlier detection. Clustering results are validated by Elbow plots and Silhouette coefficient, details of which can be studied in [49].

Clustering methods can be categorized broadly into partitioning methods, density based, model based and search based methods etc., discussed in following section [50].

A. Partitioning Methods:

Partitioning methods use distance-based matrices for clustering and partition the features into subsets [47]. This matrices work on the similarity of any unsupervised evaluation criteria of features. This method produces non-overlapping spherical shaped clusters after one level partitioning. Partitioning methods are subdivided into three types: relocation based, grid based, subspace clustering.

Relocation Based Partitioning Algorithms:

These algorithms works on relocalization of clusters around a randomly selected feature. These methods use distance based metrices. Relocalization of clusters can be done either (i) around centre of gravity (k-mean or k-centroid) of a subset of features, including outliers, or (ii) around the descriptor located near the centre (k-medoid) of a subset of features, excluding outliers [51]. Process is iterated till the perfect round or near round

ISSN- 2394-5125 VOL 7, ISSUE 19, 2020

clusters are obtained. k-means algorithm is one of the oldest, standard, popular and simplest clustering algorithms [52]. The Linde-Buzo-Gray (LBG) algorithm works on same k-means, which was suggested for signal compression using vector quantization (VQ) (Nag 2019). Although k- Means clustering is still one of the most popular clustering algorithms yet it suffers from few limitations such as absence of any universal method for partitioning or clustering of firsts set of clusters and high sensitivity to outliers and noise. It also includes even quite far away located descriptor in cluster and thus distorts the shape of cluster. Due to its limitation, k-mean algorithm has been subjected to many modified versions which are ISODATA, Forgy, bisecting k-means, x-means and kernel k-means. Partitioning method based clustering algorithms that work on k-medoids method are PAM (Partitioning around medoid) [53], CLARA (Clustering for large applications) [4] and CLARANS [54]. Fuzzy c-means (FCM) is the clustering method which make overlapping clusters [55].

Grid Based Partitioning:

In these algorithms given data is divided into number of equal size grids (quantizes space). All the processes of clustering are performed in these grids. Grid based methods are faster because instead of processing large number of features, they process a finite number of features in a grid, hence computational complexity is reduced. STING is an example of grid-based clustering [56].

Subspace-Clustering Based Partitioning:

These methods are suitable for high dimensional data. In these methods subspaces are extracted from high dimensional data, automatically, and then clustering is performed in those subspaces. This allows better clustering compared to clustering of original multidimensional space. In this method different clustering method can be applied for different subspace. CLIQUE [56], MAFIA [57], ENCLUS , PROCLUS and ORCLUS [58]are the examples of these types of algorithms.

Density Based Clustering Methods

These methods make clusters around centres of high density areas of subset of selected data. If all sub sets of data are located around one point then round, concentric clusters are formed and if the densities of subsets of data are not localized around a centre, but are scattered, then irregular shaped clusters like: S-shaped, C-shaped are formed. Data points in dense regions will form a cluster while data points from different clusters will be separated by low density regions. Some examples of density based clustering algorithms are DBSCAN (density-based spatial clustering of applications with noise) [59], AUTOCLASS [60], SNOB[61]), MCLUST[62], EM (Expectation-maximization) [63] and OPTICS [64] are some examples of density-based clustering methods.

B. Hierarchical clustering or Connectivity based clustering

Instead of representing a cluster as round, ovoid, C or S shape, these methods represent clusters in form of a tree called as dendrogram [65]. This is hard clustering approach. Agglomerative (bottom-up) and divisive (top-down) are the two approaches used for hierarchical clustering. Hierarchical clustering methods can be based upon single linkage, complete linkage and average linkage. SLINK [66] and CLINK are the example of bottom-up approaches, while DIANA (Divisive analysis) [53], AGNES (Agglomerative nesting), UPGMA (Unweighted pair group method with arithmetic mean) [67], CLARANS (Clustering large applications based upon randomized Search) and MONA [53] are the example of top-down approaches. The major limitation of hierarchical clustering is that once a descriptor is included in a cluster, it cannot be included in other clusters of hierarchy, hence some improved hierarchical algorithms such as BIRCH (Balanced iterative reducing and clustering using hierarchies) [68], CURE (Clustering using representatives) [69], ROCK [70] and CHAMELEON [71] were also developed.

C. Model Based Clustering Methods

These methods cluster the data upon a given mathematical model. 'Decision Trees' and 'Neural Networks' are two most frequently used model based clustering methods.

Decision Tree (DT)

Decision trees [72] are used for clustering and classification. DTs follow the 'hierarchical trees model' for clustering and classification. DT has two components; decision nodes and leaves. Decision nodes are the point where data split and leaves are the decision taken at nodes along with some information about decision. DTs are easy to built, train, interpret and explain. DTs can classify both categorical and numerical variables, and are able to predict high degree of non-linearity in relationship between predictor and response. Being prone to over-fitting is the main disadvantage of DTs, which can be reduced by using a suitable pruning method.

COBWEB [73] is one of the best examples of DT based algorithms, used for nominal variables but is not suitable for clustering of big data. Other examples of DT based clustering algorithms are CLUSTER/2 [74], GALOIS [75], GCF [76], ITERATE [77], LABYRINTH [78], SUBDUE [79], UNIMEM [80] and WITT [81].

ISSN- 2394-5125 VOL 7, ISSUE 19, 2020

Neural Networks or Artificial Neural Network (ANN)

Artificial Neural Network (also known as universal function approximators) [82] is a network of artificial neurons, simulating natural neuron network of brain. Basic unit of an ANN is a perceptron, consisted on input and output layers of neurons only. A perceptron can deal with linear functions only. In order to deal with non-linear relations a multilayer network consisting of an input layer, hidden layer and output layer is used [83]. In an ANN neurons of all layers are connected to each other by neuronal connections called synapses, and like human brain, learning is performed by changing the strength of the synaptic connections. Input is received by input layer neurons and, weightage (w) and bias (b) related to each input signal is processed in neurons. Input neurons calculate both 'w' and 'b' related to it and, according to the result and a pre-set activation function (threshold), it is decided whether the input signal should be fired or activated. After receiving further inputs the neuron transmit the information downstream to other connected neurons in a process called 'forward pass' and at the then information is passed to output layer. The output of test data is compared with output of learning model and if there is any error, it is sent backward for error minimization. A perceptron is similar to support vector machine (SVM) [84] in dealing with linear function approximation, but in SVM function approximation is achieved by use of maximum margin optimization while in ANN it is achieved by incremental learning algorithms, like least square error optimization methods. 'Self-organizing map' (SOM) [85] is a popular ANN algorithm used for vector quantization, feature selections and clustering. Advantages of ANN are that they can model functional relationships in a highly nonlinear data. ANN take account of all the features for final decision making, a quality which is absent in decision tree. ANN are they cannot work efficiently with multiple target categories and provides no information about how a learning model was built, hence work like a black box.

D. Variants of Clustering Methods

Many variations of clustering methods arising after hybridization, ensemble or modification of existing methods have been proposed by different authors for the purpose of specific applications. Due to complexities they can't be categorized in either of the above groups and being discussed under variants of clustering methods separately.

Evolutionary approaches and Natural phenomenon based clustering methods

Evolutionary approaches [86] follow the concept of a natural phenomenon like evolution, pollination, swarming of bees, pollination of flowers, movement of ants etc. In evolution, many genotypes (attributes) are present in natural population (dataset), during their life time many random mutations and recombinations take place in genotype of a natural population. All these mutations and recombinations affect the fitness (survival capacity) of the natural population and at the end only the fittest one survives for next generation (subset) of population. This concept of evolution is adopted in ES (Evolution strategies) [87], EP (Evolutionary programming) [88], GA (Genetic algorithms) [89] algorithms. Besides some algorithms works on natural phenomena of living being, PSO (Particle swarm optimization) [90] is based on bees swarming or natural Brownian movement of particles in space, ACO (Ant colony optimization) is based on movement of ants in tunnel [91], FPA (Flower pollination algorithm) and its variations [92] are based on pollination. PSO, ACO, CFPA and SMO (Spider monkey algorithms) [93]; are some other popular 'evolutionary approaches' based algorithms. All these approaches use parallel stochastic search techniques. SA (Simulated annealing) based algorithms is a mixed 'sequential and parallel' stochastic search technique [94], designed to avoid bad solution and focus on the solutions which correspond to local optima of the objective functions.

Collaborative Fuzzy Clustering

Collaborative fuzzy clustering [95] is a robust clustering algorithm, capable of processing very large datasets, which are distributed on several sites like data from different databases. Main feature of CFC is the ability of processing data from different sources together and revealing the common structures which are present in more than on databases or datasets. CFC engages separate clustering algorithms for each datasets, which collaborate with each other to find the common structures in those datasets. All clustering algorithms performing under CFC establish communication link with each other and form partition matrices. These partition matrices are formed at the level of fuzzy groups instead of actual groups present in databases. There are mainly two forms of collaborative clustering; horizontal and vertical collaborative clustering [96]. In horizontal collaborative clustering, same database is split into different subsets of features, each subset having all patterns in the database. This type of clustering has been applied for Mamdani type fuzzy inference system [97], which is used to find association between datasets. In vertical collaborative clustering, database is divided into subsets of patterns such that each pattern of any subset has all features. Other than structure understandings CFC is also used to find security issues.

Graph (Theoretic) Clustering:

Graph-theoretical clustering methods [98] are robust in finding the difference. These algorithms consider each data as nodes (vertices) in a graph space and each node is separated from other at a fixed or varying distance. A

complete graph is drawn by connecting each node with all its neighbouring nodes by means of edges. Graph theoretical clustering is process of making groups of nodes taking into consideration that there should be maximum possible nodes with each cluster leaving as minimum possible nodes outside the clusters or unclustered. A well-known graph-theoretic algorithm is based on the MST (Minimal spanning tree) [99]. Spectral Clustering, proposed by Donath and Hoffman [100], is an emerging technique under graph clustering which uses eigenvectors of matrices derived from the data.

V. CLASSIFICATION:

Classification categorizes all the features of a datasets into mostly two (yes/no) or more (three to five like, yes/ may be yes/ no) categories, unlike fuzzy clustering it doesn't keep an attribute into mare than one cluster [101]. Different data classification algorithms are divided into three types [102], characteristic of each of them are as follows:

A. Discriminant or Discriminator Analysis

This type of analysis finds a clear discriminating feature amongst the given features for classification. Linear and Quadratic are the two main types of discriminant analysis used for classification.

Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) [103] is a method of classification as well as dimensionality reduction in which all the discriminating features of a class are considered and preserved. It separates two or more classes by straight lines. The feature of preserving the discriminating features of the classes make it superior to PCA [104] in which few discriminating features of classes are lost. LDA is used when the dataset to be classified has multiple variables and each variable is to be studied independently from each other. In addition, variables must be distinctly categorized as independent (predictors) and one dependent (criteria) variable. All independent variables in LDA must be normally distributed and measured as continuous values. If independent variables are measured categorically then instead of LDA, 'linear correspondence analysis' (LCA) is used. LDA combines two or more independent variable and forms a new linear variable, which is used further for classification. This resulting combination of variables is further used for reduction of dimensionality and computational complexities. Fisher's Linear Discriminant analysis [105] is one of the popular LDA algorithms used for dimensionality reduction and classification. Advantage of LDA is that it performs classification and dimensionality reduction at same time. Disadvantages of LDA are that it can't be used for unsupervised data and if independent variables are not normally distributed.

Quadratic Discriminant analysis (QDA)

Quadratic discriminant analysis (QDA) [106] separates two or more classes of response by a quadric surface (curved lines), instead of a straight lines of LDA. Unlike LDA, covariance of each class is different in QDA and must be measured separately. It is particularly useful when covariance of each class is distinct and already known. Due to estimation of separate covariance for each class, QDA has greater number of effective parameter for classification than LDA. A disadvantage of QDA is that it cannot be used as a dimensionality reduction technique and it must be used with care when the feature space is large. A discriminant method that compromise between LDA and QDA is Regularized Discriminant Analysis (RDA) [107]. In RDA coordinate axes are rotated in such a way to maximize inter-class and intra-class variance among the classes.

B. Probabilistic methods:

These methods categorized the features on the basis of their probability of keeping to a particular class. These methods are divided into following subtypes:

Logistic Regression

When the response of a feature is measured as numerical (quantitative) data then logistic regression is used [108]. Amongst other type of regression, this is the only type of regression used for classification. Linear regression methods are one to one interaction prediction methods. These methods model the linear relationship between independent variable (predictor) and dependent variables (criteria), to predict the value of response after change in value of predictor. Logistic regression methods are used to model nonlinear relationship when there are more than one predictors (X_1,X_2,X_3…….X_n) and insert their combined effect on one (Y). Logistic regression has wide application in market research to predict the customer's choice, which is usually dependent upon many factors of products (like quality, pack size, use and cost, etc.), Cancer microarray data is also same type of datasets where final effect (cancer) depends upon combined outcome of many attributes (genes). Logistic regression calculates the results based on probabilities of the logistic curve rather than normal (bell shaped) curve, and does not require homogeneity of variance. It works well with numerical and categorical variables, both. But it can't predict the interrelationship among different predictors because it assumes that predictors are independent to each other, which is possible in DT and ANN.

ISSN- 2394-5125 VOL 7, ISSUE 19, 2020

Naive-Bayesian Classifiers

Classification of big-data can be performed using statistical methods or computer based machine learning approaches, both of which demands much time and computational cost for data with high dimensionality. In addition, these approaches focus on one attribute at a time for classification. Bayes proposed a method that follow human thinking pattern (use feature similarities or previous decisions) to classify multidimensional data. The drawback of Bayesian prediction is that no two human thinks in the same way, hence they would classify the same data in different ways. Naïve Bayesian classifiers (NBCs) [109] overcome this limitation and use two common assumptions for classification of any datasets. First assumption follows Fisher's approach, which considers the similarity of a predictor of the dataset with others. Second assumption follows Bayesian approach which consider the previous history (decision) of classification used for similar kind of data. NBCs deal one attribute (dimension) at a time and classify a new attribute based on 'joint posterior probability', which is the sum of 'prior probability' and 'likelihood'. They consider classification is process of judgment call. Both the approaches are combined in NBCs because shedding either of approach could not provide a suitable classification of multidimensional data. Advantages of NBCs are that they are suitable for large number of variables, fast to be trained and give results consequently. NBCs are insensitive to unimportant variables. Main disadvantage of NBCs is that like logistic regression they also cannot be predicted interaction among different variables.

C. Model based classifiers

Support Vector Machines (SVM)

SVM is a distance based classification method. It is a non-probabilistic binary linear classifier, which separates the given data in two distinct binary classes separated from each other by a clear gap, using its internal learning algorithms. SVM algorithms use labelled training data (supervised learning), to give output in form of an optimal hyper-plane which categorizes new examples on either of its side. Hyper-plane divides the data into two distinct classes where one class in on one side and other on second side of line. In two dimensional spaces this hyper-plane is a line, while in three dimensional spaces it is a plane surface [110]. SVM classifiers are good for labelled data, but also suitable for unlabelled categorical data too. SVM classifiers and NBCs incorporates wrapper FS with them.

k-Nearest-Neighbours (k-NN or KNN) Classifiers

Nearest Neighbour methods are non-parametric learning algorithms, which don't need any prior assumption about distribution of data from where the training datasets are drawn. Training of both positive and negative attributes is involved in them. After training model building, new instances are classified by calculating the distance to the nearest training case called neighbour and the sign of that point then determines the classification of the sample. k-NN classifier [109] algorithms requires finite number (k) of neighbours for classifying a test sample. k can be equal to or lesser than the number of feature present in training dataset. If k-value is high it eliminates the noise of data but increases the computational complexity, while a smaller k-value increase the effect of noise.

Advantages of k-NN are that it is easy to implement and give good results if feature are weighted proportionately. Limitations of k-NN include high sensitivity to irrelevant parameters and drastic negative impact on classification accuracy if outliers or irrelevant variables, having undue weightage are included in classification. In addition, k-NN is slow if training set has many examples.

Decision tree (DT)

Basic feature of DT are given in section IV. E. For classification, DT follow 'divide-and-conquer' algorithm and being a supervised method used a training data set for further to classification of data. There are two main types of DT: Classification trees and Regression trees. In classification trees the decision variables are categorical (discrete) and such trees are built through binary recursive partitioning. An iterative process of splitting the data into partitions and then splitting it further at each of the nodes, till are the features are exhausted, is used to build classification trees. In regression tress decision variables are continuous in nature. DT are not suitable for classification of datasets with multiple classes. Classification and regression trees (CART) [111], GALOIS, GCF, ITERATE, LABYRINTH, SUBDUE, UNIMEM and WITT are some example of DT based algorithms.

Random forest (RF)

RF classifier [112] is a supervised ensemble learning algorithm, which combines more than one algorithms of same or different kind for classification. A RF is a meta-classification approach that fits a number of sub classifiers (DTs) on various subsets of a dataset and average from each DT is used to improve the accuracy of classification. RF is a model made up of many DTs and is easiest and the most flexible algorithm for classification. It makes DTs from randomly selected datasets and over-fit many DTs to make forest. Prediction about best feature class is calculated from each DTs and the best solution (class) among all them is selected on the basis of voting from each DTs. RFs. Boruta algorithm [113], an algorithm for selecting important feature

ISSN- 2394-5125 VOL 7, ISSUE 19, 2020

works on RFs. Advantages of RFs are that they reduce the over-fitting and variance, thus improves accuracy. RFs run efficiently or large databases and can handle thousand of variables in data without deleting any variables thus gives equal importance to all the variables. RFs can be used to predict missing data accurately and don't need a separate test for error estimation. Disadvantages of RF are that they require high computational power and resources, and longer time for training and processing. Furthermore RF is like a black box where programmer has no control over internal working of model.

Artificial Neural Networks (ANN)

Besides clustering of unsupervised data, ANN is also used for classification of supervised data [114]. Details of ANN are described earlier in section IV.F.

D. Variants of Classification Algorithms

Learning vector quantization (LVQ)

Neural networks are the best machine learning methods to deal with very complex data, but they require big datasets, lot of computational power and much processing time. When data size is complex but small and limited computational power is available, simple and intuitive machine learning methods like LVQ [115] can be used for classification. LVQ is a non-parametric method which support both binary and multiclass classification problems and works on a competitive (winner-take-all, Hebbian learning-based approach) learning strategy. It is a process of classifying the patterns where each output unit represents a class. In LVQ a primary random pool of vectors is prepared which are then applied for training of model. LVQ are it is simple, intuitive, and easy to implement while still yielding decent performance. SOM [116] algorithm that is a similar algorithm to LVQ but for unsupervised learning. Advantages of LVQ over ANN are that unlike ANN LVQ is not a black box and we can learn how the learning instances look like. LVQ is a baseline technique that was defined with a few variants LVQ1, LVQ2, LVQ2.1, LVQ3, OLVQ1, and OLVQ3 as well as many third-party extensions and refinements too numerous to list. Disadvantages of LVQ are that the learning rate is typically linearly decayed over the training period from an initial value to close to zero. The more complex the class distribution, the more codebook vectors that will be required, some problems may need thousands.

CHAID(Chi-square automatic interaction detector)

When a categorical response has multiple outcomes instead of binary, then CHAID [117] is used for multi-way splits from a single parent node, instead of CART. CHAID makes many multi-way frequency tables and that's why very useful for market researches and for classification of data having multiple outcomes of response. CHAID is a good algorithm for initial steps of classification owe to being fast. It builds wide DTs for better representation of data. Disadvantage of CHAID is that it requires big data volume to get dependable results because it treats the data into fragments. Second disadvantage of CHAID it that instead of building binary tree it make irregular shaped trees which looks unrealistic and some time not have pleasant appearance.

VI. APPLICATION OF BIG DATA MINING

Applications of big-data are not limited to the field of computer science, communication science, information sciences only. It is a multidisciplinary science, which is mainly used for network and communication, networking, network security. Besides the other fields where big-data mining is very useful are medical science, basic and applied researches in science and biology, climate change, library science and management, marketing, finance business and crime data analysis etc. These many areas where big-data mining is used are briefly described below.

A. Basic computer science research:

All the applied aspects of computer science such as developing softwares, algorithms, customized package for customers etc require application of various aspects of big-data mining. Without applying big-data mining methods, applied computer science researches would be compromised.

B. Information and communication technology

Present age is also known as the age of Information and technology. Managing of huge number of transmissions, building LAN, MAN, WAN etc, their proper functioning and protection from viruses and other spyware need a thorough understanding and application of big-data mining technologies.

C. Biological Science:

Large volume of gene, protein sequence, image data, mass spectrometry data, genomics, proteomics data, wet laboratory data, pharmacological data and clinical trial data are the various types of biological data generated on daily basis. Biomarkers, drug targets, new drug discovery etc are made possible by use of computer science and data mining technologies. Genomics, proteomics, next generation sequencing, bioinformatics, real-time PCR,

ISSN- 2394-5125 VOL 7, ISSUE 19, 2020

microarray are some area where high dimensional and high volume data is generated. Microarray data is multidimensional data consisting of rows and columns, which all together makes 1000s of data instances together. Microarray is used to study the simultaneous expression of 1000s of genes together from different biological samples. These sample can be disease Vs normal, samples from different developmental stage etc., depending upon which the microarray data can be clinical or non clinical. Comparison of gene expression profile and its data mining using various data mining algorithms helps in identification of disease biomarkers, disease/development related gene networks etc. K-mean, Genetic algorithms, ANN[114], DT [118], RF [119], SVM [110] are the few common algorithms used for study of microarray data.

D. Environmental Science:

Environment is a vast term that includes everything surrounding us. Impact of environment on life forms, especially on human life, are vice-versa, favorable environment support the growth of life forms while positive and negative activities of human affects the environment in same way. Although researches in environmental science and climate are going on for a long time, use of computer science in these areas has stated for almost two decades. At present vast data related to status and conservation of forest, water and other natural resources, status of green house gases, pollutions, rainfall, etc is being generated, which requires big-data analytics for its proper study. Salih et al [120] used data mining for prediction of sediment being carried by river. Their study was focused on effect of sediment deposition on big dams and reservoirs, and devising new machines to alleviate this problem. Ma et al [121] and Lin and Xiao [122] studied global water system and marine water system respectively for forecasting the change in these ecosystems. Both group argued that proper forecasting will help in future management of national aquatic resources. Su et al [123] showed the use of big data mining for study of global carbon emission pattern and climate change. They advocated that in future data mining based modelling and studies must be adopted for indept studies of climate change data. Urban planning is an important parameter for environment protection of a city. Cao et al [124] highlighted importance of big data analysis for dealing with problem of traffic, noise pollution, garbage treatment and other problems of city to make it smart city. Works by Song et al [125] and Faghmous and Kumar [126] discussed various theoretical and practical aspects of big data analytics for evaluation of environment performance. Use of data mining for agricultural data [127], ground water data [128], for geospatial and climate change data [129] and spatial data [130] are the example of application of data mining in various field of climate change and environment.

E. Finance and Market Analysis:

Finance, business, banking and marketing are the area, where big data analytics is very much useful. What a person, like to purchase, what are his/her earnings, expenditure and savings etc seems very small data at individual level. But then the data from a population is combined it becomes a huge, multidimensional data. Financial status, trend of purchasing, market investment, prediction of investment and gain, prediction of share market, customer buying behaviour and profitability, are the few example where very important information are obtained using data mining.

A detailed analysis of market analysis of 1034 publications from 2015-19 was performed by Lopez et al [131]. Another text mining study was performed by Pejic et al [132] to study the various aspects of Big Data Analysis in Financial Sector. A book 'Data Management, Analytics and Innovation" by Yafooz et al [133] explain thoroughly about big data analytics for business and marketing. Use of Expectation-Maximization (EM) and K-Means++ clustering algorithms on consumer buying behavior was studied by Yoseph et al [134], and based on their outcome they devised a method which increase sales growth rate from 5% to 9%. Hasan et al [135] reviewed the effect of big data studies on finance and revealed its significantly positive effect. Wright et al [136] also advocated the use of data mining in business to business B2B marketing. Nakagawa et al [137] use k-medoids clustering for predicting the trend of stock market. They used dynamic time warping with k-medoids in their work.

F. Library management:

Libraries are the place where thousands of books are kept, so that user can read the book of their interest. Traditionally, libraries contain hard copy of reading materials (books, novels, newspapers, journals etc), but in modern world libraries are also being digitalized. Digitization of reading material is creating a lot of data in electronic form. This digital would have not additional advantages over its hard form of data until it could not be accessed by users easily. In addition the digitized material can be transported in any part of globe easily via internet. Be it digital or traditional form of library, both need special data mining softwares to their management. Use of data mining for better library management and services to the readers was described by Ball [138]. Process of bibliomining process, with emphasis on data warehousing and privacy was describe in the

ISSN- 2394-5125 VOL 7, ISSUE 19, 2020

work by Nicholson [139]. Benefits of data mining for library science were mentioned by Jie [140]. Use of collaborative fuzzy clustering for University library management was demonstrated by Liu [141]. A case study of data mining in analysis of fund raising response, sent to the library user by library were done by Dole and Hill [142]. Review published by Li et al [143] highlights the technical problems of digital library and use of data mining is library services. Kubek [144] in his book on 'Centroid-Based Library Management and Document Clustering' described the application of hierarchic and centroid-based document management algorithms in library management.

G. Population data analysis:

Census data is an important data for planning development and welfare plants by governments. Census data is huge data and use of computational methods is becoming useful for deeper understanding of computational data. Jagtap [145] used WEKA tool for study of district level census, socio-economic and population related other data for knowledge discovery. Use of data mining for better e-Governance has been shown by Gupta et al. [118]. In which they showed that various administrative procedure for transport, municipal records, education, healthcare, ports and shipping, disaster management, crime and criminal tracking system and public distribution system have been made faster and smoother with use of data mining technologies.

H. Crime analysis:

Crime is negative aspect of a society. An ideal society is considered to be free from all sort of crime and criminals. Generally a crime is identified only after it has occurred and society has no mean to prevent it, all it can do is to catch the culprit and punish it after investigation. Most of the time the investigation also proves to be very complicated. Hence for a better society we need to develop the methods which would assist in accelerated investigations of crimes and prevention of crimes before it happens. Use of data mining and analytics could be very useful in crime investigation and prevention. A book by McCue [146] explained the use of data mining methods for crime analysis and prevention. Works by Chauhan and Sehgal [147], Bodare et al [148] and Vijayrani et al. [149] describe various algorithms and data mining methods used for crime analysis. Zhao and Tang [150] used data mining for urban crime analysis for proper safety, security and development of cities. Cyber crime is a new trend which is spreading with development of information technologies and digitalization. Ganesan and Mavilvahana [151] and Farsi et al [8] showed the use of data mining for cyber crime analysis. Lavanyaa and Akila [152] worked on crime against women at Tamilnadu and suggested a methodology for the same. Feng et al [153] used k-mean clustering and Naïve-Bayesian approach for study of criminal data collected from San Francisco, Chicago and Philadelphia. Use of k-mean clustering [154], Fuzzy C-mean [155], Formal Concept Analysis [156], Graph model [157] and ARIMA model[158] are example of few algorithms used for crime data analysis.

VII. CONCLUSION:

A survey of methods used during 2019-20 for microarray data mining indicates that all the available FS, clustering and classification algorithms have their inherent advantages and limitations. Hence for better results (for example prediction of disease biomarkers) multiple methods could be ensemble together, but this increases the computational cost and time. Another method to improve accuracy, reduce computational complexities and processing time could be minor or major modification of existing algorithms, and use of that modified algorithms for data mining. The altogether different approach could be hybridization of algorithms or hybrid methods. Various hybrid algorithms; hybrid Evolutionary algorithms, hybrid wrappers, hybrid filters were noticed in our survey; hence hybrid or modified methods could be a good choice. I has been proved that FS increases accuracy of classification and clustering, both, hence any of the suitable FS must be used inevitably. Computational scientists have also good scope for developing new algorithms requiring less computational complexities and processing time. Because of increasing health issues (cancer, diabetes, etc.) and improvement of microarray technologies, a huge amount of microarray data is being generated world-wide and a biologist is unable to process that data. Hence better, faster and accurate data mining algorithms are still required.

VIII. FUTURE SCOPE

According to www.forbes.com a 2.5 quintillion of data was being produced per day in 2018 and according to the website https://seedscientific.com/ **44 zettabytes of data** is estimated to exist in digital universe at the beginning of 2020. This high volume of data require quick and accurate processing to extract useful information out of it. Delay in information extraction from data would delay the policy decision that could not be taken without

proper data analysis; hence big-data mining technologies would have a huge demand in coming feature irrespective of the areas.

ACKNOWLEDGEMENTS

Authors are thankful to Department of Computer Science and Engineering for providing the guidance during this work. Authors are also obliged to the Head of the Department Prof. Y. P. Raiwani and other faculty members of the department for critical evaluation of the manuscript.

Conflict of Interest: None to declare .

IX. References:

[1]. Aggarwal, C. C., & Zhai, C. (2012). A survey of text clustering algorithms. In Mining text data (pp. 77-128): Springer.
[2]. Akter, R., & Chung, Y. (2013). An evolutionary approach for document clustering. IERI Procedia, 4(0), 370-375.
[3]. Al-Mshanji, A. A., & Al-Rashid, S. Z. (2019). Improving clustering algorithm for gene expression data using hybrid algorithm. Compusoft, 8(9), 3422-3430.
[4]. Alam, S., Dobbie, G., Koh, Y. S., Riddle, P., & Rehman, S. U. (2014). Research on particle swarm optimization based clustering: a systematic review of literature and techniques. Swarm and Evolutionary Computation, 17, 1-13.
[5]. Alyasseri, Z. A. A., Khader, A. T., Al-Betar, M. A., Awadallah, M. A., & Yang, X.-S. (2018). Variants of the flower pollination algorithm: a review. In Nature-Inspired Algorithms and Applied Optimization (pp. 91-118): Springer.
[6]. Amato, F., Lapez, A., Pea-Mandez, E. M., Vaahara, P., Hampl, A., & Havel, J. (2013). Artificial neural networks in medical diagnosis. 11(2), 47-58.
[7]. Arifando, R., Yulianto, F., Mahmudy, W. F., & Sander, B. A. (2019). Hybrid Genetic Algorithm & Learning Vector Quantization for Classification of Social Assistance Recipients. Paper presented at the 2019 International Conference on Sustainable Information Engineering and Technology (SIET).
[8]. Babu, G. P., & Murty, M. N. (1994). Clustering with evolution strategies. Pattern recognition, 27(2), 321-329.
[9]. Bagirov, A. M., Karmitsa, N., & Taheri, S. (2020). Introduction to Clustering. In Partitional Clustering via Nonsmooth Optimization (pp. 3-13): Springer.
[10]. Baliarsingh, S. K., Vipsita, S., & Dash, B. (2019). A new optimal gene selection approach for cancer classification using enhanced Jaya-based forest optimization algorithm. Neural Computing and Applications, 1-18.
[11]. Ball, R. (2019). Big data and their impact on libraries. American Journal of Information Science and Technology, 3(1), 1-9.
[12]. Biswas, G., Weinberg, J. B., & Fisher, D. H. (1998). ITERATE: A conceptual clustering algorithm for data mining. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 28(2), 219-230.
[13]. Bodare, S., Kurkute, S., Akash, M., & Pawar, R. (2019). Crime Analysis using Data Mining and Data Analytics.
[14]. Bonissone, P., Cadenas, J. M., Garrido, M. C., & Daaz-Valladares, R. A. (2010). A fuzzy random forest. International Journal of Approximate Reasoning, 51(7), 729-747.
[15]. Bracco, A., Falasca, F., Nenes, A., Fountalis, I., & Dovrolis, C. (2018). Advancing climate science with knowledge-discovery through data mining. npj Climate and Atmospheric Science, 1(1), 1-6.
[16]. Cai, Z., & Zhu, W. (2018). Multi-label feature selection via feature manifold learning and sparsity regularization. International Journal of Machine Learning and Cybernetics, 9(8), 1321-1334.
[17]. Cao, X., Wang, M., & Liu, X. (2020). Application of Big Data Visualization in Urban Planning. E&ES, 440(4), 042066.
[18]. Carpineto, C., & Romano, G. (1993). Galois: An order-theoretic approach to conceptual clustering. Paper presented at the Proceedings of ICML.
[19]. Chauhan, C., & Sehgal, S. (2017). A review: Crime analysis using data mining techniques and algorithms. Paper presented at the 2017 International Conference on Computing, Communication and Automation (ICCCA).
[20]. Cook, D. J., & Holder, L. B. (2001). Graph-Based Hierarchical Conceptual Clustering, in. Paper presented at the Journal of Machine Learning Research.

[21]. Dash, M., & Liu, H. (2003). Consistency-based search in feature selection. Artificial intelligence, 151(1-2), 155-176.
[22]. Dawyndt, P., De Meyer, H., & De Baets, B. (2006). UPGMA clustering revisited: A weight-driven approach to transitive approximation. International Journal of Approximate Reasoning, 42(3), 174-191.
[23]. Devi, B., Kumar, S., & Shankar, V. G. (2019). AnaData: A novel approach for data analytics using random forest tree and SVM. In Computing, Communication and Signal Processing (pp. 511-521): Springer.
[24]. Devi, R. V., & Sathya, S. S. (2017). Monkey behavior based algorithms-a survey. International Journal of Intelligent Systems and Applications, 9(12), 67.
[25]. Do, C. B., & Batzoglou, S. (2008). What is the expectation maximization algorithm? Nature biotechnology, 26(8), 897-899.
[26]. Dole, W. V., & Hill, J. B. (2017). Community Users in Academic Libraries: Data-Mining for Fund-Raising. Qualitative and Quantitative Methods in Libraries, 105-110.
[27]. Ester, M., Kriegel, H.-P., Sander, J. r., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. Paper presented at the Kdd.
[28]. Faghmous, J. H., & Kumar, V. (2014). A big data guide to understanding climate change: The case for theory-guided data science. Big data, 2(3), 155-163.
[29]. Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A. Y. (2014). A survey of clustering algorithms for big data: Taxonomy and empirical analysis. IEEE transactions on emerging topics in computing, 2(3), 267-279.
[30]. Farsi, M., Daneshkhah, A., Far, A. H., Chatrabgoun, O., & Montasari, R. (2018). Crime data mining, threat analysis and prediction. In Cyber Criminology (pp. 183-202): Springer.
[31]. Feng, M., Zheng, J., Han, Y., Ren, J., & Liu, Q. (2018). Big data analytics and mining for crime data analysis, visualization and prediction. Paper presented at the International Conference on Brain Inspired Cognitive Systems.
[32]. Fisher, D. (1995). Optimization and Simplification of Hierarchical Clusterings. Paper presented at the KDD.
[33]. Gabriel, R., Hoppe, T., & Pastwa, A. (2010). Classification of Metadata Categories in Data Warehousing-A Generic Approach. Paper presented at the AMCIS.
[34]. Ganesan, M., & Mayilvahanan, P. (2017). Cyber Crime Analysis in Social Media Using Data Mining Technique. International Journal of Pure and Applied Mathematics, 116(22), 413-424.
[35]. Ghosh, M., Guha, R., Sarkar, R., & Abraham, A. (2019). A wrapper-filter feature selection technique based on ant colony optimization. Neural Computing and Applications, 1-19.
[36]. Gorade, S. M., Deo, A., & Purohit, P. (2017). A study of some data mining classification techniques. International Research J. of Engineering and Technology (IRJET), 4.
[37]. Guha, S., Rastogi, R., & Shim, K. (1998). CURE: an efficient clustering algorithm for large databases. ACM Sigmod record, 27(2), 73-84.
[38]. Guha, S., Rastogi, R., & Shim, K. (2000). ROCK: A robust clustering algorithm for categorical attributes. Information systems, 25(5), 345-366.
[39]. Gunter, S., & Bunke, H. (2002). Self-organizing map for clustering in the graph domain. Pattern Recognition Letters, 23(4), 405-417.
[40]. Gupta, B., Rawat, A., Jain, A., Arora, A., & Dhami, N. (2017). Analysis of various decision tree algorithms for classification in data mining. Int. J. Comput. Appl, 163(8), 15-19.
[41]. Hajipour, M., Etminani, K., Rahmatinezhad, Z., Soltani, M., Etemad, K., Eslami, S. (2019). A Predictive Model for Mortality of Patients with Thalassemia using Logistic Regression Model and Genetic Algorithm. International Journal of Health Studies, 4(3).
[42]. Hamid, H., Zainon, F., & Tan, P. Y. (2016). Performance analysis: an integration of principal component analysis and linear discriminant analysis for a very large number of measured variables. Research Journal of Applied Sciences, 11(11), 1422-1426.
[43]. Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques: Elsevier.
[44]. Han, K., Venable, R. M., Bryant, A.-M., Legacy, C. J., Shen, R., Li, H. (2018). Graph-Theoretic Analysis of Monomethyl Phosphate Clustering in Ionic Solutions. The Journal of Physical Chemistry B, 122(4), 1484-1494.
[45]. Hasan, M. M., Popp, J. z., & OlÃ¡h, J. (2020). Current landscape and influence of big data on finance. Journal of Big Data, 7(1), 1-17.
[46]. He, X., Cai, D., & Niyogi, P. (2006). Laplacian score for feature selection. Paper presented at the Advances in neural information processing systems.
[47]. Hoque, N., Bhattacharyya, D. K., & Kalita, J. K. (2014). MIFS-ND: A mutual information-based feature selection method. Expert Systems with Applications, 41(14), 6371-6385.

[48]. Huang, J. Z., Xu, J., Ng, M., & Ye, Y. (2008). Weighting method for feature selection in k-means. Computational Methods of feature selection, 193-209.
[49]. Idri, A., Benhar, H., FernÃ¡ndez-AlemÃ¡n, J. L., & Kadi, I. (2018). A systematic map of medical data preprocessing in knowledge discovery. Computer methods and programs in biomedicine, 162, 69-85.
[50]. Inkaya, T., Kayaligil, S., & Ozdemirel, N. E. (2015). Ant colony optimization based clustering methodology. Applied Soft Computing, 28, 301-311.
[51]. Islam, K., & Raza, A. (2020). Forecasting Crime Using ARIMA Model. arXiv preprint arXiv:2003.08006.
[52]. Jadhav, S. D., & Channe, H. P. (2016). Comparative study of K-NN, naive Bayes and decision tree classification techniques. International Journal of Science and Research (IJSR), 5(1), 1842-1845.
[53]. Jagtap, S. B. (2013). Census data mining and data analysis using WEKA. arXiv preprint arXiv:1310.4647.
[54]. Jiang, B., Wang, X., & Leng, C. (2018). A direct approach for sparse quadratic discriminant analysis. The Journal of Machine Learning Research, 19(1), 1098-1134.
[55]. Jie, C. (2016). Analysis of Data Mining in the Service of the University Library. Journal of Academic Library and Information Science, 34(2), 53-57.
[56]. Joshi, A., Sabitha, A. S., & Choudhury, T. (2017). Crime analysis using K-means clustering. Paper presented at the 2017 3rd International Conference on Computational Intelligence and Networks (CINE).
[57]. Jothi, R., Mohanty, S. K., & Ojha, A. (2018). Fast approximate minimum spanning tree based clustering algorithm. Neurocomputing, 272, 542-557.
[58]. Jovic, A., Brkic, K., & Bogunovic, N. (2015). A review of feature selection methods with applications. Paper presented at the 2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO).
[59]. Kalbkhani, H., Salimi, A., & Shayesteh, M. G. (2015). Classification of brain MRI using multi-cluster feature selection and KNN classifier. Paper presented at the 2015 23rd Iranian Conference on Electrical Engineering.
[60]. Kannan, S. S., & Ramaraj, N. (2010). A novel hybrid feature selection via Symmetrical Uncertainty ranking based local memetic search algorithm. Knowledge-Based Systems, 23(6), 580-585.
[61]. Kapoor, P., Singh, P. K., & Cherukuri, A. K. (2020). Crime data set analysis using formal concept analysis (FCA): A survey. In Advances in Data Sciences, Security and Applications (pp. 15-31): Springer.
[62]. Karypis, G., Han, E.-H., & Kumar, V. (1999). Chameleon: Hierarchical clustering using dynamic modeling. Computer, 32(8), 68-75.
[63]. Kaufman, L., & Rousseeuw, P. J. (2009). Finding groups in data: an introduction to cluster analysis (Vol. 344): John Wiley & Sons.
[64]. Kaur, M., & Kang, S. (2016). Market Basket Analysis: Identify the changing trends of market data using association rule mining. Procedia computer science, 85, 78-85.
[65]. Kerr, G., Ruskin, H. J., Crane, M., & Doolan, P. (2008). Techniques for clustering gene expression data. Computers in biology and medicine, 38(3), 283-293.
[66]. Kesavaraj, G., & Sukumaran, S. (2013). A study on classification techniques in data mining. Paper presented at the 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT).
[67]. Khan, S. I., & Hoque, A. S. M. L. (2015). Development of national health data warehouse for data mining. Database Systems Journal, 6(1), 3-13.
[68]. Kharis, S. A. A., Hadi, I., & Hasanah, K. A. (2019). Multiclass Classification of Brain Cancer with Multiple Multiclass Artificial Bee Colony Feature Selection and Support Vector Machine. Paper presented at the Journal of Physics: Conference Series.
[69]. Kim, S. (2020). A miRNA-and mRNA-seq-Based Feature Selection Approach for Kidney Cancer Biomakers. Cancer Informatics, 19, 1176935120908301.
[70]. Kirubha, V., & Priya, S. M. (2016). Survey on data mining algorithms in disease prediction. Int J Comput Trends Tech, 38(3), 24-128.
[71]. Kononenko, I. (1994). Estimating attributes: analysis and extensions of RELIEF. Paper presented at the European conference on machine learning.
[72]. Kriegel, H.-P., Kruger, P., & Zimek, A. (2009). Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. ACM Transactions on Knowledge Discovery from Data (TKDD), 3(1), 1-58.
[73]. Krovi, R. (1992). Genetic algorithms for clustering: a preliminary investigation. Paper presented at the Proceedings of the Twenty-Fifth Hawaii International Conference on System Sciences.
[74]. Kubek, M. (2020). Centroid-Based Library Management and Document Clustering. In Concepts and Methods for a Librarian of the Web (pp. 103-116): Springer.
[75]. Kumar, V., & Minz, S. (2014). Feature selection: a literature review. SmartCR, 4(3), 211-229.

[76]. Kursa, M. B., Jankowski, A., & Rudnicki, W. R. (2010). Boruta-a system for feature selection. Fundamenta Informaticae, 101(4), 271-285.
[77]. Ladha, L., & Deepa, T. (2011). Feature selection methods and algorithms. International journal on computer science and engineering, 3(5), 1787-1797.
[78]. Lavanyaa, S., & Akila, D. (2019). Crime against Women (CAW) Analysis and Prediction in Tamilnadu Police Using Data Mining Techniques. International Journal of Recent Technology and Engineering (IJRTE), 7(5C).
[79]. Lebowitz, M. (1987). Experiments with incremental concept formation: Unimem. Machine learning, 2(2), 103-138.
[80]. Lee, S., Hyun, Y., & Lee, M.-J. (2019). Groundwater potential mapping using data mining models of big data analysis in Goyang-si, South Korea. Sustainability, 11(6), 1678.
[81]. Lei, T., Jia, X., Zhang, Y., He, L., Meng, H., & Nandi, A. K. (2018). Significantly fast and robust fuzzy c-means clustering algorithm based on morphological reconstruction and membership filtering. IEEE Transactions on Fuzzy Systems, 26(5), 3027-3041.
[82]. Li, C., & Biswas, G. (1997). Unsupervised clustering with mixed numeric and nominal data-a new similarity based agglomerative system. Paper presented at the International Workshop on AI and Statistics.
[83]. Li, S., Jiao, F., Zhang, Y., & Xu, X. (2019). Problems and changes in digital libraries in the age of big data from the perspective of user services. The Journal of Academic Librarianship, 45(1), 22-30.
[84]. Lin, B., & Xiao, F. (2020). Application Engineering of Big Data Technology for Global Marine Environment Forecasting.
[85]. Liu, Y. (2018). Data mining of university library management based on improved collaborative filtering association rules algorithm. Wireless Personal Communications, 102(4), 3781-3790.
[86]. Lokers, R., Knapen, R., Janssen, S., van Randen, Y., & Jansen, J. (2016). Analysis of Big Data technologies for use in agro-environmental science. Environmental Modelling & Software, 84, 494-504.
[87]. Lopez-Robles, J. R., Rodriguez-Salvador, M., Gamboa-Rosales, N. K., Ramirez-Rosales, S., & Cobo, M. J. (2019). The last five years of Big Data Research in Economics, Econometrics and Finance: Identification and conceptual analysis. Procedia computer science, 162, 729-736.
[88]. Ma, H., Xiong, Y., Hou, X., & Shu, Q. (2020). Application of Big Data in Water Ecological Environment Monitoring. MS&E, 750(1), 012044.
[89]. Madzarov, G., Gjorgjevikj, D., & Chorbev, I. (2009). A multi-class SVM classifier utilizing binary decision tree. Informatica, 33(2).
[90]. Mak, M.-W., & Kung, S.-Y. (2008). Fusion of feature selection methods for pairwise scoring SVM. Neurocomputing, 71(16-18), 3104-3113.
[91]. Makinde, O. S. (2019). Gene expression data classification: some distance-based methods. Kuwait Journal of Science, 46(3).
[92]. Markey, M. K., Lo, J. Y., Tourassi, G. D., & Floyd Jr, C. E. (2003). Self-organizing map for cluster analysis of a breast cancer database. Artificial Intelligence in Medicine, 27(2), 113-127.
[93]. McCue, C. (2014). Data mining and predictive analysis: Intelligence gathering and crime analysis: Butterworth-Heinemann.
[94]. Miao, J., & Niu, L. (2016). A survey on feature selection. Procedia Computer Science, 91, 919-926.
[95]. Miner, G., Elder Iv, J., Fast, A., Hill, T., Nisbet, R., & Delen, D. (2012). Practical text mining and statistical analysis for non-structured text data applications: Academic Press.
[96]. Mohammed, M. A., & Al-Khafaji, H. (2017). Maximal itemsets mining algorithm based on Bees' Algorithm. Paper presented at the 2017 Annual Conference on New Trends in Information & Communications Technology Applications (NTICT).
[97]. Mohapatra, S. K., & Mohanty, M. N. (2020). Big Data Analysis and Classification of Biomedical Signal Using Random Forest Algorithm. In New Paradigm in Decision Science and Management (pp. 217-224): Springer.
[98]. Momenzadeh, M., Sehhati, M., & Rabbani, H. (2019). A novel feature selection method for microarray data classification based on hidden Markov model. Journal of biomedical informatics, 95, 103213.
[99]. Nakagawa, K., Imamura, M., & Yoshida, K. (2019). Stock price prediction using kâ€ • medoids clustering with indexing dynamic time warping. Electronics and Communications in Japan, 102(2), 3-8.
[100]. Nascimento, M. C. V., & De Carvalho, A. C. (2011). Spectral methods for graph clustering: a survey. European Journal of Operational Research, 211(2), 221-231.
[101]. Nayyar, A., & Puri, V. (2017). Comprehensive analysis & performance comparison of clustering algorithms for big data. Review of Computer Engineering Research, 4(2), 54-80.
[102]. Ng, R. T., & Han, J. (2002). CLARANS: A method for clustering objects for spatial data mining. IEEE transactions on knowledge and data engineering, 14(5), 1003-1016.

[103]. Nicholson, S. (2003). The Bibliomining Process: Data Warehousing and Data Mining for Library Decision-Making. In.

[104]. Oyelade, J., Isewon, I., Oladipupo, F., Aromolaran, O., Uwoghiren, E., Ameh, F. (2016). Clustering algorithms: their application to gene expression data. Bioinformatics and Biology insights, 10, BBI. S38316.

[105]. Pawar, P. M., Nielsen, R. H., Prasad, N. R., Ohmori, S., & Prasad, R. (2012). Gcf: Green conflict free tdma scheduling for wireless sensor network. Paper presented at the 2012 IEEE international conference on communications (ICC).

[106]. Pejic Bach, M., Krstic, Z., Seljan, S., & Turulja, L. (2019). Text mining for big data analysis in financial sector: A literature review. Sustainability, 11(5), 1277.

[107]. Peng, C., Wu, X., Yuan, W., Zhang, X., & Li, Y. (2019). MGRFE: multilayer recursive feature elimination based on an embedded genetic algorithm for cancer classification. IEEE/ACM transactions on computational biology and bioinformatics.

[108]. Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on pattern analysis and machine intelligence, 27(8), 1226-1238.

[109]. Raval, U. R., & Jani, C. (2016). Implementing & improvisation of K-means clustering algorithm. International Journal of Computer Science and Mobile Computing, 5(5), 191-203.

[110]. Ristoski, P., & Paulheim, H. (2016). Semantic Web in data mining and knowledge discovery: A comprehensive survey. Journal of Web Semantics, 36, 1-22.

[111]. Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., Amancio, D. R., Costa, L. d. F. (2019). Clustering algorithms: A comparative approach. PloS one, 14(1).

[112]. Sajana, T., Rani, C. M. S., & Narayana, K. V. (2016). A survey on clustering techniques for big data mining. Indian journal of Science and Technology, 9(3), 1-12.

[113]. Salazar, H., Gallego, R. n., & Romero, R. n. (2006). Artificial neural networks and clustering techniques applied in the reconfiguration of distribution systems. IEEE Transactions on Power Delivery, 21(3), 1735-1742.

[114]. Salih, S. Q., Sharafati, A., Khosravi, K., Faris, H., Kisi, O., Tao, H. (2020). River suspended sediment load prediction based on river discharge information: application of newly developed data mining models. Hydrological Sciences Journal, 65(4), 624-637.

[115]. Sanchez-Torrubia, G., & Torres-Blanc, C. (1993). A Mamdani-type fuzzy inference system to automatically assess Dijkstraâ€™s algorithm simulation. INFORMATION THEORIES & APPLICATIONS, 35.

[116]. Saritas, M. M., & Yasar, A. (2019). Performance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification. International Journal of Intelligent Systems and Applications in Engineering, 7(2), 88-91.

[117]. Seifollahi, S., Bagirov, A., Zare Borzeshi, E., & Piccardi, M. (2019). A simulated annealing • based maximum • margin clustering algorithm. Computational Intelligence, 35(1), 23-41.

[118]. Senliol, B., Gulgezen, G., Yu, L., & Cataltepe, Z. (2008). Fast Correlation Based Filter (FCBF) with a different search strategy. Paper presented at the 2008 23rd international symposium on computer and information sciences.

[119]. Shen, Y., & Pedrycz, W. (2017). Collaborative fuzzy clustering algorithm: Some refinements. International Journal of Approximate Reasoning, 86, 41-61.

[120]. Shirazi, F., & Rashedi, E. (2016). Detection of cancer tumors in mammography images using support vector machine and mixed gravitational search algorithm. Paper presented at the 2016 1st conference on swarm intelligence and evolutionary computation (CSIEC).

[121]. Shukla, A. K., Singh, P., & Vardhan, M. (2019). A new hybrid wrapper TLBO and SA with SVM approach for gene expression data. Information Sciences, 503, 238-254.

[122]. Shukla, A. K., Singh, P., & Vardhan, M. (2020). Gene selection for cancer types classification using novel hybrid metaheuristics approach. Swarm and Evolutionary Computation, 54, 100661.

[123]. Siswantining, T., Kamelia, T., & Sarwinda, D. (2018). Implementation of Chi Square Automatic Interaction Detection (CHAID) Method to Identify Type 2 Diabetes Mellitus in Tuberculosis Patient. A Case Study in Cipto Mangunkusumo Hospital. Paper presented at the Journal of Physics: Conference Series.

[124]. Song, M.-L., Fisher, R., Wang, J.-L., & Cui, L.-B. (2018). Environmental performance evaluation with big data: Theories and methods. Annals of Operations Research, 270(1-2), 459-472.

[125]. Soni, N., & Ganatra, A. (2012). Categorization of several clustering algorithms from different perspective: a review. International Journal of.

[126]. Sreejith, A. G., Lansy, A., Krishna, K. S. A., Haran, V. J., & Rakhee, M. (2020). Crime Analysis and Prediction Using Graph Mining. In Inventive Communication and Computational Technologies (pp. 699-705): Springer.
[127]. Stepp, R. E., & Michalski, R. S. (1986). Conceptual clustering of structured objects: A goal-oriented approach. Artificial Intelligence, 28(1), 43-69.
[128]. Su, Y., Yu, Y., & Zhang, N. (2020). Carbon emissions and environmental management based on Big Data and Streaming Data: A bibliometric analysis. Science of The Total Environment, 138984.
[129]. Sublime, J., Grozavu, N., Cabanes, G., Bennani, Y., & Cornuejols, A. (2015). From horizontal to vertical collaborative clustering using generative topographic maps. International journal of hybrid intelligent systems, 12(4), 245-256.
[130]. Suebsing, A., & Hiransakolwong, N. (2009). Feature selection using euclidean distance and cosine similarity for intrusion detection model. Paper presented at the 2009 First Asian Conference on Intelligent Information and Database Systems.
[131]. Talmon, J. L., Fonteijn, H., & Braspenning, P. J. (1993). An analysis of the WITT algorithm. Machine Learning, 11(1), 91-104.
[132]. Tang, J., Alelyani, S., & Liu, H. (2014). Feature selection for classification: A review. Data classification: Algorithms and applications, 37.
[133]. Tavakkol, B., Jeong, M. K., & Albin, S. L. (2019). Measures of Scatter and Fisher Discriminant Analysis for Uncertain Data. IEEE Transactions on Systems, Man, and Cybernetics: Systems.
[134]. Tharwat, A. (2016). Linear vs. quadratic discriminant analysis classifier: a tutorial. International Journal of Applied Pattern Recognition, 3(2), 145-180.
[135]. Theodoridis, S., & Koutroumbas, K. (2009). clustering algorithms ii: Hierarchical algorithms. In Pattern Recognition (Fourth Edition): Academic Press.
[136]. Thompson, K., & Langley, P. (1991). Concept formation in structured domains. In Concept Formation (pp. 127-161): Elsevier.
[137]. Timofeev, R. (2004). Classification and regression trees (CART) theory and applications. Humboldt University, Berlin, 1-40.
[138]. Tran, T. N., Drab, K., & Daszykowski, M. (2013). Revised DBSCAN algorithm to cluster data with dense adjacent clusters. Chemometrics and Intelligent Laboratory Systems, 120, 92-96.
[139]. Trelles, O., Prins, P., Snir, M., & Jansen, R. C. (2011). Big data, but are we ready? Nature Reviews Genetics, 12(3), 224-224.
[140]. Tseng, L. Y., & Yang, S. B. (1997). Genetic algorithms for clustering, feature selection and classification. Paper presented at the Proceedings of International Conference on Neural Networks (ICNN'97).
[141]. Verdiguel, N., Feng, Z., Westbrook, J., & Zardecki, C. (2019). Data Mining Scientific Literature Demonstrates Use of Biological and Medical Data Across Scientific Disciplines. The FASEB Journal, 33(1_supplement), 493.410-493.410.
[142]. Vijayarani, S., Suganya, E., & Navya, C. (2020). A Comprehensive Analysis of Crime Analysis Using Data Mining Techniques.
[143]. Vishwakarma, S., Nair, P. S., & Rao, D. S. (2017). A Comparative Study of K-means and K-medoid Clustering for Social Media Text Mining. INTERNATIONAL JOURNAL, 2(11).
[144]. Wang, S., & Yuan, H. (2014). Spatial data mining: a perspective of big data. International Journal of Data Warehousing and Mining (IJDWM), 10(4), 50-70.
[145]. Witten, I. H., & Frank, E. (2002). Data mining: practical machine learning tools and techniques with Java implementations. Acm Sigmod Record, 31(1), 76-77.
[146]. Wolf, L., & Shashua, A. (2005). Feature selection for unsupervised and supervised inference: The emergence of sparsity in a weight-based approach. Journal of Machine Learning Research, 6, 1855-1887.
[147]. Wright, L. T., Robin, R., Stone, M., & Aravopoulou, D. E. (2019). Adoption of Big Data technology for innovation in B2B marketing. Journal of Business-to-Business Marketing, 26(3-4), 281-293.
[148]. Xiao, Q.-L., Zheng, H., & Yao, Q.-A. (2019). Feature Selection for Cancer Classification Based on SRPSO Algorithm. DEStech Transactions on Engineering and Technology Research(icicr).
[149]. Xuan, G., Zhu, X., Chai, P., Zhang, Z., Shi, Y. Q., & Fu, D. (2006). Feature selection based on the bhattacharyya distance. Paper presented at the 18th International Conference on Pattern Recognition (ICPR'06).
[150]. Yafooz, W. M. S., Bakar, Z. B. A., Fahad, S. K. A., & Mithon, A. M. (2020). Business Intelligence Through Big Data Analytics, Data Mining and Machine Learning. In Data Management, Analytics and Innovation (pp. 217-230): Springer.
[151]. Yamini, M. P. C. (2019). A violent crime analysis using fuzzy c-means clustering approach. ICTACT Journal on Soft Computing, 9(3), 1939-1944.

ISSN- 2394-5125 VOL 7, ISSUE 19, 2020

[152]. Yang, C., Yu, M., Hu, F., Jiang, Y., & Li, Y. (2017). Utilizing cloud computing to address big geospatial data challenges. Computers, Environment and Urban Systems, 61, 120-128.
[153]. Yang, P., Liu, W., Zhou, B. B., Chawla, S., & Zomaya, A. Y. (2013). Ensemble-based wrapper methods for feature selection and class imbalance learning. Paper presented at the Pacific-Asia Conference on Knowledge Discovery and Data Mining.
[154]. Yoseph, F., Ahamed Hassain Malim, N. H., HeikkilÃ¤, M., Brezulianu, A., Geman, O., & Paskhal Rostam, N. A. (2020). The impact of big data market segmentation using data mining and clustering techniques. Journal of Intelligent & Fuzzy Systems(Preprint), 1-15.
[155]. Zhang, D. (2019). Support vector machine. In Fundamentals of Image Data Mining (pp. 179-205): Springer.
[156]. Zhang, T., Ramakrishnan, R., & Livny, M. (1996). BIRCH: an efficient data clustering method for very large databases. ACM Sigmod Record, 25(2), 103-114.
[157]. Zhao, X., & Tang, J. (2018). Crime in urban areas: A data mining perspective. ACM SIGKDD Explorations Newsletter, 20(1), 1-12.
[158]. Zhao, Z. A., & Liu, H. (2011). Spectral Feature Selection for Data Mining (Open Access): CRC Press.